What is claimed is:

1. A method comprising:

reading one or more corpus, each of the one or more corpus comprising a plurality of clinical trial descriptions;

reading a list of disease conditions, the list having one or more disease condition mapped to one or more associated keyword;

determining a plurality of categories of clinical trials based on the plurality of clinical trial descriptions and the list of disease conditions;

determining medical criteria from each of the plurality of clinical trial descriptions, wherein the medical criteria comprise inclusion criteria and/or exclusion criteria and wherein determining medical criteria comprises applying an artificial neural network to the plurality of clinical trial descriptions;

determining a frequency of occurrence for each of one or more repeated terms from each category of clinical trials;

determining a set of category-specific repeated terms having a frequency of occurrence greater than a predetermined threshold;

determining whether each repeated term in the set of category-specific repeated terms is not present in a predetermined medical taxonomy to thereby identify a set of new terms;

determining a plurality of vectors for each new term in the set of new terms, each vector for a particular new term corresponding to the one or more associated keyword;

based on the plurality of vectors, selecting one or more of the new terms in the set of new terms; and

mapping each of the selected new terms to a disease condition thereby generating a supplemented list of disease conditions.

~~4.~~2. The method of claim ~~2~~1, wherein determining medical criteria comprises applying an artificial neural network to the plurality of clinical trial descriptions.

~~5.~~3. The method of claim ~~2~~1, further comprising:

reading a plurality of patient medical profiles;

determining one or more likely disease conditions for each patient medical profile based on the list of disease conditions;

determining one or more relevant patient medical profiles based on the determined medical criteria and the plurality of patient medical profiles;

for the relevant patient medical profiles, selecting one or more category of the plurality of categories of clinical trials based on the one or more likely disease conditions of the respective patient medical profile.

~~6.~~4. The method of claim ~~2~~1, further comprising:

reading a plurality of patient medical profiles;

determining one or more likely disease conditions for each patient medical profile based on the supplemented list of disease conditions;

determining one or more relevant patient medical profiles based on the determined medical criteria and the plurality of patient medical profiles;

for the relevant patient medical profiles, selecting one or more category of the

plurality of categories of clinical trials based on the one or more likely disease conditions of the respective patient medical profile.

~~7.~~5. The method of claim 1, wherein reading one or more corpus comprises accessing the one or more corpus via an application programming interface (API).

~~8.~~6. The method of claim 1, wherein the one or more keywords are unique for each disease condition.

~~9.~~7. The method of claim 1, wherein each of the plurality of categories of clinical trials corresponds to a unique disease condition.

~~10.~~8. The method of claim 1, wherein determining a frequency of occurrence comprises determining a score for each repeated term.

~~11.~~9. The method of claim 8, wherein the score represents a ratio of frequency of occurrence of the repeated word in its respective category of clinical trial to the frequency of occurrence in all other categories of clinical trials.

~~12.~~10. The method of claim 1, wherein the predetermined threshold is based on frequency of occurrence in a known medical taxonomy.

~~13.~~11. The method of claim 1, further comprising, for each new term in the set of new terms, determining a probability that the new term is a medical term.

~~14.~~12. The method of claim 1, further comprising determining a condition metric, wherein the plurality of distances are determined based in part on the condition metric.

~~15.~~13. The method of claim 12, wherein the condition metric is determined from a Map/Reduce cluster.

~~16.~~14. The method of claim 1, wherein each vector in the plurality of vectors comprises a frequency the particular new term and associated keyword appear together in the respective category of clinical trial, the frequency they appear in proximity to a third medical term, and the morphological resemblance between the particular new term and associated keyword.

~~17.~~15. The method of claim 14, wherein morphological resemblance is scored based on a frequency analysis of the morphological structures and the number of their appearance in the corpus.

~~18.~~16. The method of claim 1, wherein selecting the one or more new terms comprises determining one or more vectors of the plurality of vectors having a vector magnitude below a vector magnitude threshold.

~~19.~~17. The method of claim 1, wherein selecting the one or more new terms comprises determining one or more vectors of the plurality of vectors having a minimum vector magnitude.

~~20.~~18. A system comprising:

a data store;

a computing node comprising a computer readable storage medium having program instructions embodied therewith, the program instructions executable by a processor of the computing node to cause the processor to perform a method comprising:

reading one or more corpus, each of the one or more corpus comprising a plurality of clinical trial descriptions;

reading a list of disease conditions from the data store, the list having one or more disease condition mapped to one or more associated keyword;

determining a plurality of categories of clinical trials based on the plurality of clinical trial descriptions and the list of disease conditions;

determining medical criteria from each of the plurality of clinical trial descriptions, wherein the medical criteria comprise inclusion criteria and/or exclusion criteria and wherein determining medical criteria comprises applying an artificial neural network to the plurality of clinical trial descriptions;

determining a frequency of occurrence for each of one or more repeated terms from each category of clinical trials;

determining a set of category-specific repeated terms having a frequency of occurrence greater than a predetermined threshold;

determining whether each repeated term in the set of category-specific repeated terms is not present in a predetermined medical taxonomy to thereby identify a set of new terms;

determining a plurality of vectors for each new term in the set of new terms, each vector for a particular new term corresponding to the one or more associated keyword;

based on the plurality of vectors, selecting one or more of the new terms in the set of new terms; and

mapping each of the selected new terms to a disease condition thereby generating a supplemented list of disease conditions in the data store.

~~21.~~19. A computer program product for supplementing a keyword mapping to disease conditions based on clinical trial descriptions, the computer program product comprising a computer readable storage medium having program instructions embodied therewith, the program instructions executable by a processor to cause the processor to perform a method comprising:

reading one or more corpus, each of the one or more corpus comprising a plurality of clinical trial descriptions;

reading a list of disease conditions, the list having one or more disease condition mapped to one or more associated keyword;

determining a plurality of categories of clinical trials based on the plurality of clinical trial descriptions and the list of disease conditions;

determining medical criteria from each of the plurality of clinical trial descriptions, wherein the medical criteria comprise inclusion criteria and/or exclusion criteria and wherein determining medical criteria comprises applying an artificial neural network to the plurality of clinical trial descriptions;

determining a frequency of occurrence for each of one or more repeated terms from each category of clinical trials;

determining a set of category-specific repeated terms having a frequency of occurrence greater than a predetermined threshold;

determining whether each repeated term in the set of category-specific repeated terms is not present in a predetermined medical taxonomy to thereby identify a set of new terms;

determining a plurality of vectors for each new term in the set of new terms, each vector for a particular new term corresponding to the one or more associated keyword;

based on the plurality of vectors, selecting one or more of the new terms in the set of new terms; and

mapping each of the selected new terms to a disease condition thereby generating a supplemented list of disease conditions.

~~22.~~20. A method comprising:

reading a plurality of categories of clinical trials, each category of clinical trials corresponding to a unique disease condition and having a plurality of associated keywords;

receiving a new medical term, wherein the new medical term is not present in any

of the plurality of categories of clinical trials;

for each category in the plurality of categories of clinical trials, comparing the new medical term to each associated keyword to determine for each new medical term and associated keyword pair:

a distance metric between the new medical term and associated keyword;

double occurrence of the new medical term and associated keyword;

triple occurrences of the new medical term, associated keyword, and an additional medical term;

determining a vector magnitude for each new medical term and associated keyword pair.

~~23.~~21. The method of claim 20, wherein the distance metric comprises morphological similarity.

~~24.~~22. The method of claim 20, wherein the distance metric comprises semantic similarity.

~~25.~~23. The method of claim 20, wherein the distance metric comprises syntactic similarity.

~~26.~~24. The method of claim 20, wherein double occurrences comprise joint occurrences in the same clinical trial.

~~27.~~25. The method of claim 20, wherein double occurrences comprise joint occurrences in the same category.

~~28.~~26. The method of claim 20, wherein double occurrences comprise closeness of the new medical term and associated keyword in the same category.

~~29.~~27. The method of claim 20, wherein triple occurrences comprise occurrences of the new medical term and associated keyword pair with an additional medical term.

~~30.~~28. The method of claim 20, wherein triple occurrences comprise a number of different additional medical terms with which the new medical term and associated keyword pair has joint occurrence.

~~31.~~29. The method of claim 20, further comprising selecting a smallest vector magnitude from the new medical term and associated keyword pairs.

~~32.~~30. A system comprising:

a data store;

a computing node comprising a computer readable storage medium having program instructions embodied therewith, the program instructions executable by a processor of the computing node to cause the processor to perform a method comprising:

reading from the datastore a plurality of categories of clinical trials, each category of clinical trials corresponding to a unique disease condition and having a plurality of associated keywords;

receiving a new medical term, wherein the new medical term is not present in any

of the plurality of categories of clinical trials;

for each category in the plurality of categories of clinical trials, comparing the new medical term to each associated keyword to determine for each new medical term and associated keyword pair:

a distance metric between the new medical term and associated keyword;

double occurrences of the new medical term and associated keyword;

triple occurrences of the new medical term, associated keyword, and an additional medical term;

determining a vector magnitude for each new medical term and associated keyword pair.

~~33.~~31. A computer program product for determining a vector between an unknown medical term and a known medical term, the computer program product comprising a computer readable storage medium having program instructions embodied therewith, the program instructions executable by a processor to cause the processor to perform a method comprising:

reading a plurality of categories of clinical trials, each category of clinical trials corresponding to a unique disease condition and having a plurality of associated keywords;

receiving a new medical term, wherein the new medical term is not present in any of the plurality of categories of clinical trials;

for each category in the plurality of categories of clinical trials, comparing the new medical term to each associated keyword to determine for each new medical term

and associated keyword pair:

a distance metric between the new medical term and associated keyword;

double occurrences of the new medical term and associated keyword;

triple occurrences of the new medical term, associated keyword, and an additional medical term;

determining a vector magnitude for each new medical term and associated keyword pair.